
Query-Aware Graph Attention for Precise Subgraph Retrieval in Knowledge-Augmented Reasoning

Anonymous Author(s)

Affiliation

Address

email

Abstract

In recent years, Retrieval-Augmented Generation (RAG) has demonstrated great potential in enhancing the factual accuracy of large language models (LLMs) in open-domain question answering. Incorporating knowledge graphs (KGs) as external knowledge sources into the RAG paradigm is a promising direction. However, KG-RAG systems for complex multi-hop reasoning tasks still face significant challenges in precisely retrieving structured evidence highly relevant to the query. Existing approaches struggle to dynamically and accurately retrieve graph-based evidence by effectively leveraging query semantics and relational information. To address these challenges, we propose a novel framework called Query-aware Subgraph Retrieval Augmented Generation (QSRAG), centered around a new attention-based architecture termed Query-Relational Graph Attention Network (QR-GAT). QR-GAT is a graph attention mechanism that learns expressive representations of triples by capturing intricate interactions between the query context and relation types. Based on these representations, a scoring module assigns fine-grained relevance scores to triples in the KG, enabling precise subgraph retrieval for downstream reasoning. These structured evidence subgraphs, enriched with confidence scores, are then provided to an LLM to enhance its reasoning capability. Extensive experiments on two widely-used multi-hop Knowledge Graph Question Answering (KGQA) datasets, WebQSP and CWQ, demonstrate that our approach achieves state-of-the-art retrieval performance, particularly excelling in identifying complex multi-hop evidence. KGQA results further show that QSRAG delivers state-of-the-art or competitive performance on both datasets. Our work highlights the effectiveness of query-aware graph attention for accurate structured evidence retrieval, and its potential to enhance knowledge-augmented reasoning with large language models.

1 Introduction

The emergence of large language models (LLMs) has significantly advanced the field of natural language processing [Brown et al., 2020, Huang and Chang, 2022, Wei et al., 2022]. However, due to the static and limited nature of their internal knowledge, LLMs often struggle with tasks that demand precise factual grounding or complex reasoning [Kasai et al., 2023, Ji et al., 2023]. Retrieval-Augmented Generation (RAG), which integrates external knowledge into LLMs, has proven to be an effective solution to this limitation [Borgeaud et al., 2022, Gao et al., 2023]. Among external knowledge sources, knowledge graphs (KGs)—structured and large-scale repositories of factual knowledge—are particularly well-suited to support RAG, especially in enhancing LLMs’ capabilities in domain-specific or multi-hop question answering [Guo et al., 2024, Pan et al., 2024, Peng et al., 2024].

Effectively integrating knowledge graphs into the RAG paradigm (KG-RAG) requires addressing a key challenge: how to accurately and efficiently retrieve structured evidence (e.g., key triples, paths, or subgraphs) that is highly relevant to a given natural language query from a complex and interconnected graph structure. Although various KG-RAG methods have been proposed, we observe that they generally face three major obstacles:

Firstly, it is still difficult to accurately identify relevant evidence in a multi-hop structure. Answers in knowledge graphs are usually embedded in lengthy multi-hop reasoning paths. Although existing retrieval methods can highlight entities and relationships on the actual reasoning path[Jiang et al., 2023a, Luo et al., 2024b, Sun et al., 2024], they may still be disturbed by irrelevant neighbors or redundant information, resulting in a lack of focus in the retrieval.

Second, insufficient modeling of interactions between query semantics and graph structure hinders performance. Many graph retrieval or GNN-based methods fail to dynamically incorporate the semantic intent of the query into the information aggregation process, instead encoding the graph independently of the query[Mavromatis and Karypis, 2024, Yasunaga et al., 2021, Li et al., 2025]. This limits their ability to adapt attention based on the specific reasoning needs of each question.

Third, retrieval noise and inefficiencies in the usage of LLM contexts present practical challenges. Due to inadequate relevance modeling, retrieved triples often include irrelevant or weakly related facts. Feeding such noisy evidence into LLMs with limited context capacity dilutes useful information[Xu et al., 2023], making it harder for the model to identify and utilize key reasoning evidence. In addition, some KG-RAG methods involve iterative LLM calls during retrieval or reasoning, which substantially increases inference latency[Jin et al., 2024, Gao et al., 2024, Kim et al., 2023, Ma et al., 2024, Xiong et al., 2024, Sun et al., 2024].

Graph neural networks (GNNs) are naturally suited for modeling multi-hop connections in structured data, making them promising tools for subgraph retrieval[Li et al., 2024, Mavromatis and Karypis, 2022]. Their ability to capture complex dependencies among nodes facilitates path discovery and structural reasoning. However, standard GNNs often fall short in query-specific and fine-grained evidence selection. Our work seeks to address this gap by building on the strengths of GNNs while introducing mechanisms for dynamic, query-aware retrieval.

To systematically tackle the above challenges, we propose a novel framework, Query-aware Subgraph Retrieval Augmented Generation (QSRAG). The core of our framework is a query-aware and relation-guided graph attention mechanism called Query-Relational Graph Attention Network (QR-GAT). This mechanism enables efficient and accurate subgraph retrieval from KGs, thereby enhancing downstream LLM reasoning[Brody et al., 2022].

QR-GAT introduces a customized attention mechanism that fuses global query semantics with explicit relation types and applies it to the message passing process of GNNs. This design enables QR-GAT to dynamically adjust the attention to entities and relations based on the input query, thereby solving the problem of insufficient interaction between query semantics and graph structure. By emphasizing entities and relations that are highly relevant to the query, QR-GAT is able to better identify and enhance information on key multi-hop reasoning paths, effectively alleviating the difficulty of evidence alignment in complex graph structures. In addition, with the help of learned node representations and fine-grained attention weights, our computational scoring module is able to accurately score triples based on their relevance to the query. Then, the top-k triples with the highest scores are selected to construct a concise and focused subgraph, which helps reduce redundancy and noise in retrieval results and improves the efficiency and effectiveness of downstream reasoning within the limited context window of LLM.

In the reasoning stage, the retrieved subgraph and its confidence scores are provided to the LLM to guide final answer generation. Notably, unlike some methods [Luo et al., 2024b,a] that require fine-tuning the LLM for specific retrieval formats or tasks, potentially compromising its generality and interpretability, our framework employs LLMs in a plug-and-play manner, leveraging high-quality structured evidence without any fine-tuning. Our key contributions are summarized as follows:

- We propose a knowledge-augmented reasoning framework, QSRAG, to address the challenge of structured evidence retrieval in LLM-based Knowledge Graph Question Answering (KGQA). It is centered on a novel graph attention mechanism, QR-GAT, designed for precise, query-aware, and relation-guided triple scoring and subgraph retrieval.

- We conduct extensive experiments on WebQSP and CWQ datasets, and the results show that QR-GAT achieves state-of-the-art results on standard retrieval metrics.
- We show that integrating subgraphs retrieved by QR-GAT significantly improves LLM reasoning performance, achieving state-of-the-art or competitive KGQA results on both datasets.
- We perform comprehensive ablation studies to validate the role of the query-relation attention mechanism in achieving accurate retrieval, and show that incorporating confidence scores positively impacts LLM reasoning, offering insights into effective evidence utilization.

The remainder of this paper is organized as follows. Section 2 reviews related work in KG-based RAG and graph neural networks. Section 3 introduces preliminaries and problem formulation. Section 4 describes our approach, including the QR-GAT architecture and the two-stage QSRAG framework. Section 5 presents experimental setup and results, followed by further analysis and ablations in Section 6. Section 7 concludes the paper. Additional QA examples and ablation results are provided in the appendix.

2 Related Work

KG-based RAG and Graph Retrieval. Retrieval-Augmented Generation has become a pivotal paradigm for enhancing the factual capabilities of large language models. Incorporating knowledge graphs as structured external knowledge sources into the RAG framework holds great promise, particularly for complex reasoning tasks. The core challenge in this direction lies in efficiently and accurately retrieving multi-hop structured evidence that is highly relevant to a natural language query from large and complex KGs. To tackle this challenge, various strategies have been proposed. For example, Reasoning on Graphs (RoG) [Luo et al., 2024b] adopts a planning-based approach to retrieve reasoning paths. SubgraphRAG [Li et al., 2025] proposes a trainable subgraph retriever that combines MLP-based scoring with structural features for retrieval. Other methods explore graph retrieval via heuristic search [Sun et al., 2024], combinatorial optimization [He et al., 2024, Hu et al., 2024], or by converting KGs into serialized textual input for LLMs [Wu et al., 2023], as well as biologically inspired memory-augmented retrieval [Gutierrez et al., 2024], non-parametric continual memory frameworks [Gutiérrez et al., 2025], and unified retriever-reasoner architectures for KGQA [Jiang et al., 2023b]. Some approaches also utilize GNNs to assist in graph encoding and retrieval, such as GNN-RAG [Mavromatis and Karypis, 2024], which leverages graph neural representations to support LLM reasoning. While these methods vary in design and purpose, most fall short when it comes to directly assigning fine-grained relevance scores to individual KG triples based on both query semantics and relation types, followed by global ranking for optimal top-k retrieval. For instance, path-based approaches like RoG and subgraph-based scoring methods like SubgraphRAG differ from our attention-based solution in how they model query-relation interactions across multi-hop structures. As a result, the retrieved evidence from these models may be less focused, especially for complex queries. This gap highlights a critical need for more effective mechanisms that can dynamically leverage query and relation information to guide triple-level scoring and filtering on the graph. Our proposed framework directly addresses this need by enhancing the precision of structured evidence retrieval through a more expressive and query-sensitive graph mechanism.

Graph Neural Networks for KGs. GNNs are powerful tools for processing graph-structured data, and have been widely and successfully applied to knowledge graphs for tasks such as representation learning, link prediction, and graph-based reasoning. GNNs learn node and relation representations by propagating and aggregating information across graph neighborhoods, effectively capturing structural dependencies. Graph Attention Networks (GATs) [Brody et al., 2022], as a prominent subclass of GNNs, further improve expressiveness by employing attention mechanisms to dynamically weigh neighboring nodes during aggregation. Several KGQA and KG-RAG approaches have leveraged GNNs to handle KG information. For instance, one line of work introduces GRAFT-Net [Sun et al., 2018], which applies a heterogeneous GNN with attention over relations and personalized propagation to perform node classification on question-specific subgraphs that integrate both KG and textual evidence. Other approaches employ GNNs as “subgraph reasoners” to process dense subgraphs retrieved from KGs, or extract shortest paths between entities to construct serialized reasoning chains [Mavromatis and Karypis, 2024]. Variants of GATs [Gao et al., 2024] have been used to explore temporal and structural dependencies within retrieved subgraphs, often combined

with textual representations. While these models demonstrate the utility of GNNs in KG-based tasks, their architectures and applications differ from the objective of fine-grained subgraph retrieval. Most of them operate on pre-filtered or candidate subgraphs, performing downstream tasks such as classification, path extraction, or encoding. Furthermore, their attention and aggregation mechanisms are not explicitly designed to be dynamically modulated by natural language queries and relation types for triple-level relevance scoring across multi-hop structures. The limitations of standard GATs in this context underscore the importance of query- and relation-aware attention for complex retrieval tasks—a capability that has been largely underexplored. Our work builds upon the solid foundation of GNNs and GATs, introducing a novel attention mechanism that explicitly integrates query semantics and relational context into triple-level scoring. The proposed QR-GAT model directly operates on KG structures and dynamically assigns query-conditioned relevance scores to each triple, enabling precise and efficient subgraph retrieval. This offers a new perspective compared to existing GNN-based methods and brings structured evidence retrieval closer to the needs of complex multi-hop KGQA.

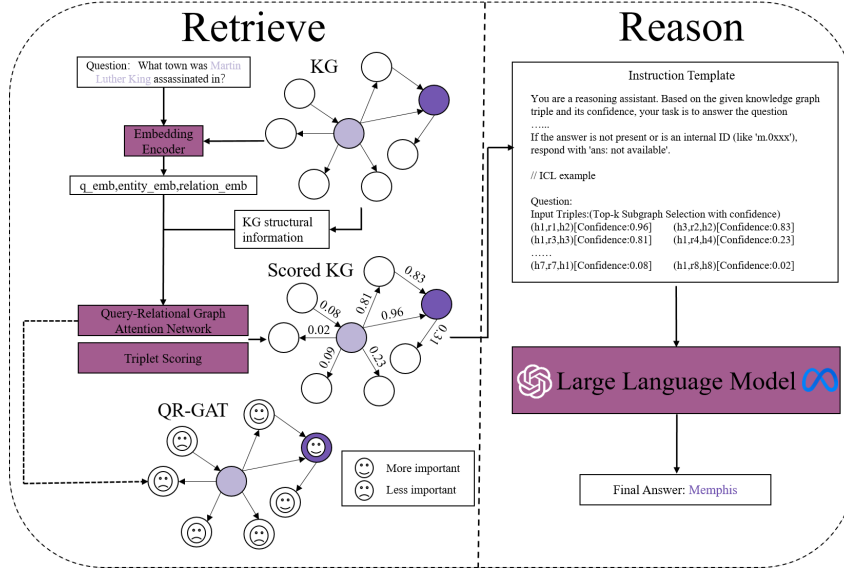


Figure 1: Overview of our QSRAG framework, consisting of (1) a Query-Relational Graph Attention Network (QR-GAT) for structured evidence retrieval from the KG, and (2) a contextual reasoning module using an LLM with in-context learning.

3 Preliminaries

In this section, we introduce key concepts relevant to our work, including knowledge graphs, knowledge graph question answering, and retrieval-augmented generation. We also formally define the problem studied in this paper.

Knowledge Graphs. A knowledge graph G is a structured representation of factual knowledge in the form of a graph. It typically consists of a set of entities E , a set of relation types R , and a set of factual triples T . Each triple $(h, r, t) \in T$ denotes a fact, where $h \in E$ is the head entity, $r \in R$ is the relation, and $t \in E$ is the tail entity.

Knowledge Graph Question Answering. KGQA refers to the task of answering a natural language question q by identifying the correct answer a from a knowledge graph G . The answer a may be a single entity or a set of entities in the graph. Solving KGQA typically requires understanding the semantic intent of the question, mapping it to the structure of the KG, and performing reasoning or structured querying to locate the correct answer.

Retrieval-Augmented Generation. RAG is a powerful paradigm that combines information retrieval with the generative capabilities of LLMs. In a typical RAG framework, given a user query, a

retrieval module first retrieves relevant knowledge pieces or evidence from an external knowledge source. These retrieved items are then concatenated with the original query and fed into a generation module—typically an LLM—to produce the final response or answer. RAG enables LLMs to incorporate external, real-time, or domain-specific knowledge, thereby reducing hallucinations and improving the factual accuracy and reliability of generation.

Problem Definition. We study complex multi-hop KGQA within the RAG framework, where the external knowledge source is a knowledge graph. Formally, given a natural language question q and a knowledge graph $G = (E, R, T)$, the goal is to train a two-stage model to output the correct answer a from G . The first stage retrieves a subgraph $S \subseteq T$ containing triples highly relevant to q that support reasoning. In the second stage, the retrieved subgraph S is provided as context to an LLM, which performs multi-hop reasoning via in-context learning (ICL)[Brown et al., 2020] to generate the final answer. These two stages work jointly to achieve accurate, interpretable KG-based QA.

Our main focus is to design an efficient and accurate retrieval module, especially to leverage the query-aware and relation-sensitive graph attention mechanism to accurately extract multi-hop evidence from large-scale knowledge graphs.

4 Methodology

We propose a retrieval-augmented generation framework enhanced by knowledge graphs to accurately retrieve structured evidence relevant to a given question and guide LLMs in reasoning. Our framework consists of two main stages: structured evidence retrieval and evidence-based contextual reasoning, as illustrated in Figure 1.

Before building the graph representation, we use the `gte-large-en-v1.5` encoder[Li et al., 2023] to obtain semantic embeddings of entities, relations, and the input question. This yields rich representations: \mathbf{e}_i for entity v_i , \mathbf{r}_{ij} for relation r_{ij} , and \mathbf{q} for the question.

4.1 Structured Evidence Retrieval

The goal of this stage is to extract a subgraph $S \subseteq T$ from the KG that contains triples most relevant to the input question. Rather than using traditional text matching or neighborhood expansion, we model the fine-grained semantic interactions among the query, entities, and relations across multi-hop paths using a novel attention mechanism—**Query-Relational Graph Attention Network (QR-GAT)**.

QR-GAT is designed to guide attention dynamically toward entities and relations that are critical for answering a specific question. It incorporates both query semantics and relation embeddings into the attention computation process. Standard GATs typically treat neighbors uniformly or based only on structure, ignoring query-specific signals. QR-GAT overcomes this limitation by introducing query- and relation-aware attention.

Each node v_i is initialized with:

$$\mathbf{x}_i^{(0)} = \text{Dropout}([\mathbf{e}_i \parallel \mathbf{q} \parallel \mathbf{p}_i])$$

where $\mathbf{e}_i \in \mathbb{R}^{d_e}$ is the entity embedding, $\mathbf{q} \in \mathbb{R}^{d_q}$ is the query embedding, and \mathbf{p}_i is a one-hot vector encoding used to label whether the entity is the topic entity.

At each layer l , we perform linear projections:

$$\mathbf{z}_i^{(l)} = W_s^{(l)} \cdot \mathbf{h}_i^{(l-1)}, \quad \mathbf{z}_j^{(l)} = W_t^{(l)} \cdot \mathbf{h}_j^{(l-1)}$$

where $W_s^{(l)}$ and $W_t^{(l)}$ are learnable weights for source and target roles respectively.

Attention score $\alpha_{ij}^{(l)}$ is computed by combining structural and query-guided terms:

$$\begin{aligned} \alpha_{ij,\text{base}}^{(l)} &= \mathbf{a}^{(l)\top} \cdot \text{LeakyReLU}(\mathbf{z}_i^{(l)} + \mathbf{z}_j^{(l)} + W_e^{(l)} \cdot \mathbf{r}_{ij}) \\ \alpha_{ij,\text{plus}}^{(l)} &= (W_q^{(l)} \cdot \mathbf{q})^\top \cdot (W_r^{(l)} \cdot \mathbf{r}_{ij}) \\ \alpha_{ij}^{(l)} &= \text{softmax}_j(\alpha_{ij,\text{base}}^{(l)} + \alpha_{ij,\text{plus}}^{(l)}) \end{aligned}$$

Node representations are updated via multi-head attention:

$$\mathbf{h}_i^{(l)} = \text{LayerNorm}([\mathbf{h}_{i,1}^{(l)} \parallel \dots \parallel \mathbf{h}_{i,H}^{(l)}])$$

where $\mathbf{h}_{i,k}^{(l)} = \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l,k)} \cdot \mathbf{z}_j^{(l)}$ for attention head k .

We use a bidirectional QR-GAT (BiQR-GAT) to encode both forward and reverse edges. Final entity representation is:

$$\mathbf{h}_i = [\mathbf{h}_i^{\rightarrow} \parallel \mathbf{h}_i^{\leftarrow}]$$

Triplet Scoring. Using final node representations \mathbf{h}_h and \mathbf{h}_t for head and tail entities, the score of a triple (h, r, t) is computed via a two-layer MLP:

$$s(h, r, t) = W_2 \cdot \text{ReLU}(W_1 \cdot [\mathbf{q} \parallel \mathbf{h}_h \parallel \mathbf{r} \parallel \mathbf{h}_t])$$

Training and Inference. The retriever is trained as a binary classifier, using a binary cross-entropy loss function to distinguish between positive triplets (triplets on the shortest path between the topic entity and the answer entity) and negative triplets. At inference time, the scores of all candidate triplets are calculated, and the top- k triplets are selected to form a structured evidence subgraph.

4.2 Evidence-based Contextual Reasoning

The retrieved top- k triples and their confidence scores are formatted into a textual prompt and fed into the LLM. We adopt an in-context learning (ICL) strategy, where the question and serialized triples are concatenated into a structured prompt of the form: (head, relation, tail) (Confidence: 0.96).

This prompt is then passed into an LLM (e.g., Llama-3.1-8B-Instruct, GPT-4o-mini, or GLM-4-Air), which generates the final answer based on both the question and the supporting evidence. Confidence scores play an important role in guiding the LLM to focus on more relevant triples, improving reasoning reliability and mitigating hallucinations.

To ensure clarity and reproducibility, we provide the full input prompt template as well as several representative multi-hop QA examples in Appendix B. These include both successful and failure cases, which showcase how different triples and confidence patterns affect the final prediction. The prompt examples demonstrate how structured knowledge is effectively integrated into the LLM’s context window, and how the LLM reasons over multi-hop paths grounded in the KG.

5 Experiments

To thoroughly evaluate the effectiveness of our proposed QSRAG (Query-aware Subgraph Retrieval Augmented Generation) framework, we conduct extensive experiments on two standard Knowledge Graph Question Answering datasets. This section details the experimental setup, evaluation results, and in-depth analysis.

5.1 Experimental Setup

Datasets. We use two widely adopted KGQA benchmarks: WebQuestionsSP (WebQSP)[Yih et al., 2016] and Complex WebQuestions (CWQ)[Talmor and Berant, 2018], both constructed over Freebase [Bollacker et al., 2008]. WebQSP comprises 4,737 questions requiring up to two-hop reasoning, reflecting relatively simple to moderately complex queries. CWQ includes 34,699 questions with higher compositionality and multi-hop requirements. We followed the preprocessing and data splits of prior work such as RoG.

Evaluation Metrics. We evaluate performance at two levels: retrieval and reasoning. For retrieval, triplet Recall@ k measures the proportion of retrieved top- k triplets that lie on the shortest path from the topic entity to the answer. Answer Recall@ k measures the proportion of answer entities covered by the subgraph formed from the retrieved top- k triplets. For KGQA, we adopt standard evaluation metrics in the KGQA field: Micro F1 measures the overall F1 performance across all question-answer pairs, better reflecting the effectiveness on common questions and questions with many answers;

Macro F1 averages the F1 for each individual question-answer pair, better reflecting the average performance of the method across different question types; Hit evaluates whether at least one of the answers generated by the model is correct; Hit@1 evaluates whether the model’s most frequently predicted answer is correct according to any of the ground truth answers.

Baselines. We compare QSRAG with several representative KGQA methods covering different technical approaches: SR+NSM w/ E2E[Zhang et al., 2022], a method combining semantic parsing and neural symbolic models, utilizing a pretrained text encoder for constrained path search. Retrieve-Rewrite-Answer[Wu et al., 2023], a multi-stage method typically involving text retrieval, question rewriting, and answer generation. RoG[Luo et al., 2024b] fine-tunes an LLM to predict relation paths and construct reasoning trajectories. G-Retriever[He et al., 2024], combining cosine similarity search with combinatorial optimization techniques, aiming to construct a subgraph capable of connecting topic entities and potential answer entities. GNN-RAG[Mavromatis and Karypis, 2024], utilizing Graph Neural Networks to predict answer entities and extracting triplets on the answer paths through structured methods like shortest path search. SubgraphRAG[Li et al., 2025] is a primary comparison object in this paper, which encodes directional structural distances as structural features using Directional Distance Encoding (DDE) and employs a lightweight multi-layer perceptron (MLP) combined with a parallel triplet scoring mechanism; its retriever design allows for the decomposition of the subgraph distribution.

Implementation Details. In terms of hardware environment, we utilize a K100-Ai cluster to complete the entire model training and inference process. For the retrieval stage, in the main experiments, we set the range for the selected number of top-k triplets, k, between 50 and 500, and conduct a detailed analysis of the impact of different k values in subsequent in-depth analysis work. In the reasoning stage, we employ multiple Large Language Models, including Llama-3.1-8B-Instruct, GPT-4o-mini-2024-07-18, and GLM-4-Air-250414. The temperature parameter for these models during inference was uniformly set to 0. In general cases, we primarily used the top 100 triplets for relevant operations, but simultaneously, we also selected representative k values, such as 50 and 200, for study and included the relevant results in the report.

In the research on the retrieval stage, we successfully reproduced the experimental results of SubgraphRAG. After verification, its metrics are consistent with those shown in their paper. Given that this method currently represents a cutting-edge performance level in the field, we directly cited the results from the SubgraphRAG paper for other experiments. In the reasoning stage experiments, we reproduced the experiments of SubgraphRAG, RoG, and GNN-RAG and obtained corresponding results. However, considering that some experiments can no longer represent the most advanced performance level in this field at present, based on strategic considerations, we focused our replication work more on those more updated and more competitive baseline methods.

5.2 Evaluation Results

Retrieval Performance. Table 1 reports the Triplet Recall@k and Answer Recall@k on WebQSP and CWQ. Our QR-GAT retriever achieves the best results on both datasets. On WebQSP, QR-GAT reaches a Triplet Recall@k of 0.905, improving over SubgraphRAG (0.883) by +2.5%, and over RoG (0.713) by a significant +26.9%. For Answer Recall@k, it achieves 0.942, nearly matching SubgraphRAG’s 0.944 and outperforming all other baselines. On the more challenging CWQ dataset, QR-GAT achieves 0.903 in Triplet Recall@k, surpassing SubgraphRAG (0.811) by +11.3%, and RoG (0.623) by +44.9%. In terms of Answer Recall@k, QR-GAT scores 0.955, improving over SubgraphRAG (0.914) by +4.5% and outperforming all others by a large margin. These results demonstrate the strong capability of our Query-Relational Graph Attention Network (QR-GAT) in accurately identifying and retrieving relevant multi-hop evidence, providing higher-quality structural input for downstream LLM reasoning.

Table 1: Retrieval evaluation results on WebQSP and CWQ datasets. Best results are in **bold**.

	WebQSP		CWQ	
	Triplet Recall	Answer Recall	Triplet Recall	Answer Recall
SR+NSM w/ E2E	0.487	0.707	–	–
Retrieve-Rewrite-Answer	0.058	0.740	–	–
RoG	0.713	0.807	0.623	0.841
G-Retriever	0.294	0.545	0.183	0.375
GNN-RAG	0.522	0.818	0.500	0.841
SubgraphRAG	0.883	0.944	0.811	0.914
QSRAG	0.905	0.942	0.903	0.955

Table 2: KGQA results on WebQSP and CWQ. Best results are in **bold**. By default, our reasoners use the top 100 retrieved triples. Results with 50 and 200 triples (indicated in parentheses) are also shown. Our best result is indicated with an underline.

	WebQSP				CWQ			
	Micro-F1	Macro-F1	Hit	Hit@1	Micro-F1	Macro-F1	Hit	Hit@1
SR+NSM w/ E2E	–	64.10	–	–	–	46.30	–	–
ToG	–	–	82.60	–	–	–	67.60	–
Retrieve-Rewrite-Answer	–	–	79.36	–	–	–	–	–
G-Retriever	–	53.41	73.46	–	–	–	–	–
RoG	52.60	70.45	85.38	79.36	46.12	54.44	60.97	56.10
GNN-RAG	10.89	71.28	85.69	80.59	28.80	59.43	66.81	61.74
SubgraphRAG + Llama (200)	42.46	66.98	83.22	77.36	39.04	41.97	54.38	48.23
SubgraphRAG + GPT-4o-mini (200)	49.78	69.76	85.54	78.89	44.82	43.00	53.33	48.12
QSRAG + Llama (50)	47.99	69.46	83.85	79.42	46.17	45.78	56.13	51.09
QSRAG + Llama	48.62	70.61	85.63	80.59	43.33	45.70	56.95	51.66
QSRAG + Llama (200)	49.03	70.13	85.55	79.54	40.82	44.55	57.18	51.03
QSRAG + GPT-4o-mini (50)	51.57	69.10	81.20	75.55	49.80	44.16	51.15	46.90
QSRAG + GPT-4o-mini	55.25	70.70	83.34	76.93	50.29	45.31	53.02	48.51
QSRAG + GPT-4o-mini (200)	55.50	71.82	84.77	78.13	49.77	45.44	54.04	48.97
QSRAG + GLM-4-Air (50)	47.47	68.69	87.44	79.06	39.74	<u>46.32</u>	63.43	52.96
QSRAG + GLM-4-Air	46.38	68.25	88.85	80.59	33.06	44.16	63.15	52.54
QSRAG + GLM-4-Air (200)	45.85	68.43	90.88	81.50	29.39	44.28	<u>65.56</u>	<u>53.46</u>

Reasoning Performance. We systematically evaluate the proposed QSRAG model on two standard question answering datasets, WebQSP and CWQ, and compare it with a series of representative methods (such as SubgraphRAG, RoG, GNN-RAG, etc.). As shown in Table 2, QSRAG performs well on WebQSP. QSRAG + GPT-4o-mini (200) achieves the current best Micro F1 (55.50) and Macro F1 (71.82) scores, which are 5.5% and 1.9% higher than RoG (Micro F1 52.60, Macro F1 70.45), and significantly outperforms SubgraphRAG + GPT-4o-mini (Micro F1 49.78, Macro F1 69.76) by 11.5% and 3.0% respectively. In terms of Hit and Hit@1 indicators, QSRAG + GLM-4-Air (200) achieves the best Hit (90.88) and Hit@1 (81.50) scores. In addition, among different large language models, Hit is always greater than 83 and Hit@1 is always greater than 75, showing consistent and strong recall capabilities.

QSRAG also achieves competitive results on the more challenging CWQ dataset. QSRAG + GPT-4o-mini achieves the best Micro F1 (50.29), 9.0% higher than RoG (46.12) and 12.2% higher than SubgraphRAG + GPT-4o-mini (44.82). GNN-RAG achieves the highest Macro F1 (59.43) among the full entries, while QSRAG + GPT-4o-mini (200) achieves a good Macro F1 (45.44), 5.7% higher than SubgraphRAG + GPT-4o-mini (43.00). In terms of Hit, ToG (67.60) leads most methods, and QSRAG + GLM-4-Air (200) obtains a comparable Hit (65.56). For Hit@1, QSRAG + GPT-4o-mini (200) (48.97) is comparable to RoG (56.10) and leads SubgraphRAG + GPT-4o-mini (48.12) by 1.8%.

On the more challenging CWQ dataset, the SubgraphRAG variants exhibited F1 scores below 45, which may suggest certain limitations in effectively handling complex graph structures compared to their performance on simpler datasets. Overall, QSRAG achieves a good balance between high recall and stable classification performance, showing strong triple reasoning ability and wide applicability.

Table 3: Impact of Retrieved Top-k on Retrieval Performance (Recall@k).

Topk	WebQSP		CWQ	
	Triple Recall	Answer Recall	Triple Recall	Answer Recall
50	0.845	0.879	0.842	0.899
100	0.900	0.919	0.903	0.931
200	0.946	0.948	0.944	0.952
300	0.963	0.959	0.963	0.962
400	0.974	0.967	0.974	0.968
500	0.980	0.971	0.979	0.972

6 Further Analysis

To gain a deeper understanding of the contributions and robustness of each component of QSRAG, we conducted several analysis experiments.

Impact of Retrieved Top-k (Retrieval). Table 3 shows the impact of different numbers of top-k triplets on QSRAG’s retrieval performance (Triplet Recall@k and Answer Recall@k) on the WebQSP and CWQ datasets on the test set. The results indicate, as shown in Table 3, that as the number of retrieved top-k triplets increases, the performance of our retriever on the Triplet Recall@k and Answer Recall@k metrics generally shows an upward trend. This suggests that expanding the retrieval scope usually captures more relevant triplets and involved entities. For example, on the WebQSP dataset, as k increases from 50 to 500, Triplet Recall increases from 0.845 to 0.980, and Answer Recall increases from 0.879 to 0.971. A similar trend is observed on the CWQ dataset, where Triplet Recall increases from 0.842 to 0.979, and Answer Recall increases from 0.899 to 0.972. This shows that increasing the number of top-k can provide more potentially relevant evidence.

Impact of Retrieved Top-k (Reasoning). Table 4 shows the impact of different numbers of top-k triplets on QSRAG’s reasoning performance (Macro-F1 and Hit) on the test set. The results indicate, as shown in Table 4, that QSRAG’s performance generally exhibits a trend of first rising and then falling as the number of top-k triplets increases. This validates that increasing relevant evidence helps improve reasoning accuracy, but there exists an optimal top-k value. Specifically, on the WebQSP dataset, both Macro-F1 and Hit metrics peak at k=100 (Macro-F1 70.61, Hit 85.63), and then gradually decrease as the number of top-k increases. On the more challenging CWQ dataset, Macro-F1 achieves its highest score at k=50 (45.78), while Hit peaks at k=200 (57.18), after which the metrics also decline. This phenomenon of performance plateauing or even slightly decreasing after increasing top-k to a certain extent contrasts with the trend of recall continuously increasing with k observed in Table 3. This strongly supports our analysis: although increasing top-k can retrieve more potentially relevant evidence, when the amount of retrieved information exceeds a certain threshold, the introduced noisy triplets may interfere with the LLM’s judgment or exceed the LLM’s capacity to effectively process the context, thereby impairing the final reasoning performance. This analysis guided our selection of top-k values in the main experiments, where we chose top-k values that performed better on the WebQSP and CWQ datasets as representatives to report.

Table 4: Impact of Retrieved Top-k on Reasoning Performance with Llama-3.1-8B-Instruct.

Topk	WebQSP		CWQ	
	Macro-F1	Hit	Macro-F1	Hit
50	69.46	83.85	45.78	56.13
100	70.61	85.63	45.70	56.95
200	70.13	85.55	44.55	57.18
300	68.80	84.89	42.70	54.46
400	67.61	83.91	41.01	52.23
500	67.58	84.83	41.64	53.19

Retriever Ablations. To assess the impact of our query- and relation-aware attention mechanism, we conduct an ablation by replacing the enhanced QR-GAT attention with a standard graph attention (i.e., removing the α_{ij}^{plus} component and retaining only α_{ij}^{base}). The results are summarized in Table 5. Compared to the full model (Plus), removing the query-relation enhancement (Base) leads to consistent performance degradation across both datasets. On WebQSP, Triplet Recall drops from 0.905 to 0.851 (-5.97%), and Answer Recall from 0.942 to 0.900 (-4.46%). On CWQ, Triplet Recall decreases from 0.903 to 0.877 (-2.88%), and Answer Recall from 0.955 to 0.940 (-1.57%). This clearly demonstrates the core contribution of our proposed query-aware and relation-guided attention mechanism in accurately identifying question-relevant triples.

Table 5: Retriever Ablation Results (Recall@k).

Model	WebQSP		CWQ	
	Triplet Recall	Answer Recall	Triplet Recall	Answer Recall
Base	0.851	0.900	0.877	0.940
Plus	0.905	0.942	0.903	0.955

7 Conclusion

In this work, we address the challenge of accurately retrieving structured evidence for knowledge graph-augmented generation by proposing the QSRAG framework, centered around the Query-Relational Graph Attention Network. QR-GAT enables precise triple scoring and subgraph retrieval through a query-aware and relation-guided attention mechanism. Experimental results demonstrate the strong retrieval performance of our method and its state-of-the-art or competitive KGQA results on both the WebQSP and CWQ datasets. These findings validate the effectiveness of precise structured evidence retrieval—guided by query-aware graph attention—in enhancing knowledge-augmented reasoning.

References

- K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008.
- S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.
- S. Brody, U. Alon, and E. Yahav. How attentive are graph attention networks? In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL <https://openreview.net/forum?id=F72ximsx7C1>.
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, and H. Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1, 2023.
- Y. Gao, L. Qiao, Z. Kan, Z. Wen, Y. He, and D. Li. Two-stage generative question answering on temporal knowledge graph using large language models. In L. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6719–6734. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.401. URL <https://doi.org/10.18653/v1/2024.findings-acl.401>.
- T. Guo, Q. Yang, C. Wang, Y. Liu, P. Li, J. Tang, D. Li, and Y. Wen. Knowledgenavigator: Leveraging large language models for enhanced reasoning over knowledge graph. *Complex & Intelligent Systems*, 10(5):7063–7076, 2024.
- B. J. Gutierrez, Y. Shu, Y. Gu, M. Yasunaga, and Y. Su. Hipporag: Neurobiologically inspired long-term memory for large language models. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract-Conference.html.
- B. J. Gutiérrez, Y. Shu, W. Qi, S. Zhou, and Y. Su. From RAG to memory: Non-parametric continual learning for large language models. *CoRR*, abs/2502.14802, 2025. doi: 10.48550/ARXIV.2502.14802. URL <https://doi.org/10.48550/arXiv.2502.14802>.
- X. He, Y. Tian, Y. Sun, N. Chawla, T. Laurent, Y. LeCun, X. Bresson, and B. Hooi. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Advances in Neural Information Processing Systems*, 37:132876–132907, 2024.
- Y. Hu, Z. Lei, Z. Zhang, B. Pan, C. Ling, and L. Zhao. GRAG: graph retrieval-augmented generation. *CoRR*, abs/2405.16506, 2024. doi: 10.48550/ARXIV.2405.16506. URL <https://doi.org/10.48550/arXiv.2405.16506>.
- J. Huang and K. C.-C. Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.
- Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.
- J. Jiang, K. Zhou, Z. Dong, K. Ye, X. Zhao, and J. Wen. Structgpt: A general framework for large language model to reason over structured data. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9237–9251. Association for Computational Linguistics, 2023a. doi: 10.18653/V1/2023.EMNLP-MAIN.574. URL <https://doi.org/10.18653/v1/2023.emnlp-main.574>.

- J. Jiang, K. Zhou, X. Zhao, and J. Wen. Unikgqa: Unified retrieval and reasoning for solving multi-hop question answering over knowledge graph. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023b. URL <https://openreview.net/forum?id=Z63RvyAZ2Vh>.
- B. Jin, C. Xie, J. Zhang, K. K. Roy, Y. Zhang, Z. Li, R. Li, X. Tang, S. Wang, Y. Meng, and J. Han. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. In L. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 163–184. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.11. URL <https://doi.org/10.18653/v1/2024.findings-acl.11>.
- J. Kasai, K. Sakaguchi, R. Le Bras, A. Asai, X. Yu, D. Radev, N. A. Smith, Y. Choi, K. Inui, et al. Realtime qa: What’s the answer right now? *Advances in neural information processing systems*, 36:49025–49043, 2023.
- J. Kim, Y. Kwon, Y. Jo, and E. Choi. KG-GPT: A general framework for reasoning on knowledge graphs using large language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 9410–9421. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.631. URL <https://doi.org/10.18653/v1/2023.findings-emnlp.631>.
- M. Li, S. Miao, and P. Li. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=JvkuZZ0407>.
- Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang. Towards general text embeddings with multi-stage contrastive learning. *CoRR*, abs/2308.03281, 2023. doi: 10.48550/ARXIV.2308.03281. URL <https://doi.org/10.48550/arXiv.2308.03281>.
- Z. Li, Q. Guo, J. Shao, L. Song, J. Bian, J. Zhang, and R. Wang. Graph neural network enhanced retrieval for question answering of llms. *CoRR*, abs/2406.06572, 2024. doi: 10.48550/ARXIV.2406.06572. URL <https://doi.org/10.48550/arXiv.2406.06572>.
- H. Luo, H. E, Z. Tang, S. Peng, Y. Guo, W. Zhang, C. Ma, G. Dong, M. Song, W. Lin, Y. Zhu, and A. T. Luu. Chatkbqa: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models. In L. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 2039–2056. Association for Computational Linguistics, 2024a. doi: 10.18653/V1/2024.FINDINGS-ACL.122. URL <https://doi.org/10.18653/v1/2024.findings-acl.122>.
- L. Luo, Y. Li, G. Haffari, and S. Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b. URL <https://openreview.net/forum?id=ZGNWW7xZ6Q>.
- S. Ma, C. Xu, X. Jiang, M. Li, H. Qu, and J. Guo. Think-on-graph 2.0: Deep and interpretable large language model reasoning with knowledge graph-guided retrieval. *CoRR*, abs/2407.10805, 2024. doi: 10.48550/ARXIV.2407.10805. URL <https://doi.org/10.48550/arXiv.2407.10805>.
- C. Mavromatis and G. Karypis. Rearev: Adaptive reasoning for question answering over knowledge graphs. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2447–2458. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-EMNLP.181. URL <https://doi.org/10.18653/v1/2022.findings-emnlp.181>.
- C. Mavromatis and G. Karypis. Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*, 2024.

- S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7): 3580–3599, 2024.
- B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, and S. Tang. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*, 2024.
- H. Sun, B. Dhingra, M. Zaheer, K. Mazaitis, R. Salakhutdinov, and W. W. Cohen. Open domain question answering using early fusion of knowledge bases and text. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4231–4242. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1455. URL <https://doi.org/10.18653/v1/d18-1455>.
- J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, L. M. Ni, H. Shum, and J. Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=nnV01PvbTv>.
- A. Talmor and J. Berant. The web as a knowledge-base for answering complex questions. In M. A. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 641–651. Association for Computational Linguistics, 2018. doi: 10.18653/V1/N18-1059. URL <https://doi.org/10.18653/v1/n18-1059>.
- J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Y. Wu, N. Hu, S. Bi, G. Qi, J. Ren, A. Xie, and W. Song. Retrieve-rewrite-answer: A kg-to-text enhanced llms framework for knowledge graph question answering. *CoRR*, abs/2309.11206, 2023. doi: 10.48550/ARXIV.2309.11206. URL <https://doi.org/10.48550/arXiv.2309.11206>.
- G. Xiong, J. Bao, and W. Zhao. Interactive-kbqa: Multi-turn interactions for knowledge base question answering with large language models. In L. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10561–10582. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.569. URL <https://doi.org/10.18653/v1/2024.acl-long.569>.
- P. Xu, W. Ping, X. Wu, L. McAfee, C. Zhu, Z. Liu, S. Subramanian, E. Bakhturina, M. Shoeybi, and B. Catanzaro. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*, 2023.
- M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec. QA-GNN: reasoning with language models and knowledge graphs for question answering. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 535–546. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.45. URL <https://doi.org/10.18653/v1/2021.naacl-main.45>.
- W.-t. Yih, M. Richardson, C. Meek, M.-W. Chang, and J. Suh. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206, 2016.
- J. Zhang, X. Zhang, J. Yu, J. Tang, J. Tang, C. Li, and H. Chen. Subgraph retrieval enhanced model for multi-hop knowledge base question answering. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5773–5784. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.396. URL <https://doi.org/10.18653/v1/2022.acl-long.396>.

A Reasoning Input Ablations.

In addition to simply removing confidence scores, we also explored various strategies to utilize these scores, particularly focusing on filtering the retrieved Top-k triplets based on confidence thresholds. Table 6 presents a comparative analysis of the impact of different confidence filtering thresholds on reasoning performance, measured by Macro-F1 and Hit metrics. The methods compared include using no confidence scores with Llama, using confidence scores with Llama, using confidence scores with GLM-4-Air, and GLM-4-Air at different confidence threshold settings.

The filtering strategy $\text{Confidence} > \text{Threshold}$ retains only triplets with scores above the specified threshold. This approach allows us to assess the effectiveness of confidence-based filtering in enhancing reasoning performance.

Table 6: Reasoning Input Ablation: Impact of Confidence Filtering Thresholds on Reasoning Performance (Macro-F1 and Hit).

Filter Strategy	WebQSP		CWQ	
	Macro-F1	Hit	Macro-F1	Hit
No Confidence + Llama	69.08	83.48	44.40	56.84
Confidence + Llama	70.61	85.63	45.70	56.95
QSRAG + GLM-4-Air	68.25	88.85	44.16	63.15
Confidence > 0.0001 + GLM-4-Air	64.09	78.77	33.94	42.98
Confidence > 0.001 + GLM-4-Air	64.18	78.43	35.63	45.02
Confidence > 0.01 + GLM-4-Air	63.37	76.17	36.77	46.58
Confidence > 0.1 + GLM-4-Air	60.11	71.13	34.40	41.55

As shown in Table 6, the method using confidence scores with Llama achieves the best performance on both datasets, with Macro-F1 and Hit metrics higher than those obtained without confidence scores. This suggests that incorporating confidence scores provides valuable signals for fine-grained evidence integration.

When applying threshold filtering with GLM-4-Air, we observe a general degradation in performance as the threshold increases. For instance, on WebQSP, increasing the threshold from 0.0001 to 0.1 results in a decrease in Macro-F1 from 64.09 to 60.11 and in Hit from 78.77 to 71.13. A similar trend is observed on CWQ.

These findings indicate that confidence scores from QR-GAT provide valuable signals to LLMs for fine-grained evidence integration. While threshold filtering may aim to remove noise, it can inadvertently discard useful information or crucial paths. Therefore, feeding all Top-k triplets along with their scores remains the most effective strategy for downstream reasoning.

B Question-answer Examples.

System :
You are a reasoning assistant. Based on the given knowledge graph triple and its confidence, your task is to answer the question. You must use only the entities found in the triplets that are ****meaningful names**** (e.g., 'Aviva Stadium', not 'm.0wz2kl3'). Each answer must be a full entity name as it appears in the triplets. Return each answer in a new line, prefixed with 'ans:'.
If the answer is not present or is an internal ID (like 'm.0xxx'), respond with 'ans: not available'.

User :
Input Triplets:
(Lou Seal,sports.mascot.team,San Francisco Giants)[Confidence: 0.9999743700027466]
(San Francisco Giants,sports.sports_team.team_mascot,Lou Seal)[Confidence: 0.9999568462371826]
(2014 World Series,sports.sports_championship_event.champion,San Francisco Giants)[Confidence: 0.8941453695297241]
(2012 World Series,sports.sports_championship_event.champion,San Francisco Giants)[Confidence: 0.8416239619255066]
(2010 World Series,sports.sports_championship_event.champion,San Francisco Giants)[Confidence: 0.8150305151939392]
(San Francisco Giants,sports.sports_team.championships,2014 World Series)[Confidence: 0.7321670055389404]
(San Francisco Giants,sports.sports_team.championships,2012 World Series)[Confidence: 0.6976805329322815]
(San Francisco Giants,sports.sports_team.championships,2010 World Series)[Confidence: 0.6664191484451294]
(San Francisco Giants,time.participant.event,2012 World Series)[Confidence: 0.07613193988800049]
(San Francisco Giants,time.participant.event,2010 World Series)[Confidence: 0.07032240927219391]
...
Question: What year did the team with mascot named Lou Seal win the World Series?

Assistant :
Based on the given triplets, we can determine that Lou Seal is the mascot of the San Francisco Giants, as indicated by the high-confidence relation (Lou Seal, sports.mascot.team, San Francisco Giants) [Confidence: 0.9999743700027466]. Furthermore, the San Francisco Giants are recorded as champions of the 2014 World Series [Confidence: 0.8941453695297241], the 2012 World Series [Confidence: 0.8416239619255066], and the 2010 World Series [Confidence: 0.8150305151939392]. Therefore, the team with mascot named Lou Seal won the World Series in the years corresponding to those championships.
ans: 2014 World Series
ans: 2012 World Series
ans: 2010 World Series

User :
Input Triplets:
(h1,r1,h2)[Confidence:0.96]
(h3,r2,h2)[Confidence:0.83]
...
Question: Which ... ?

Please provide your answer in the same format as the example. The answer should be a meaningful entity or information, not a random string like 'm02h98gq' or 'm.0xxx'.Please list all reasonable answers from the above triplets. If the answer is not present in the triplets, respond with 'ans: not available'.

Figure 2: Input prompts for KGQA.

CWQ-WebQTest-576_01e2da60a2779c4ae4b5d1547499a4f8
<p>Question: Which nation has the Alta Verapaz Department and is in Central America?</p> <hr/> <p>Retrieved Triples: Input Triplets: (Guatemala, location.country.administrative_divisions, Alta Verapaz Department)[Confidence:0.43120062351226807] (Guatemala, location.location.containedby, Central America)[Confidence: 0.4033266603946686] (Guatemala, base.aareas.schema.administrative_area.administrative_children, Alta Verapaz Department)[Confidence:0.393011152744] (Guatemala, common.topic.notable_types, Country)[Confidence: 0.30294546484947205] (Alta Verapaz Department, location.administrative_division.country, Guatemala)[Confidence: 0.16839228570461273] (Central America, location.location.contains, Costa Rica)[Confidence: 0.14474257826805115] (Central America, location.location.contains, Guatemala)[Confidence: 0.13684657216072083] ... Question: Which nation has the Alta Verapaz Department and is in Central America?</p> <hr/> <p>LLM Response: Based on the given triplets, we can determine the following: 1. Guatemala is an administrative division that contains Alta Verapaz Department. 2. Guatemala is also located in Central America. Therefore, the nation that has the Alta Verapaz Department and is in Central America is Guatemala.The formatted answer is: ans: Guatemala</p> <hr/> <p>Ground-truth: Guatemala</p>

Figure 3: Correct question-answer example 1.

CWQ-WebQTest-100_524908899a8aa334a18a0ac00f8f2fe6
<p>Question: People from the country that contains Nord-Est Department speak what languages today?</p> <hr/> <p>Retrieved Triples: Input Triplets: (French, language.human_language.countries_spoken_in, Haiti)[Confidence: 0.8332462310791016] (Haiti, base.aareas.schema.administrative_area.administrative_children, Nord-Est Department)[Confidence: 0.8295478224754333] (Nord-Est Department, location.location.containedby, Haiti)[Confidence: 0.7992613315582275] (Haiti, location.country.first_level_divisions, Nord-Est Department)[Confidence: 0.7959402203559875] (Haitian Creole, language.human_language.countries_spoken_in, Haiti)[Confidence: 0.7403706908226013] (Haiti, location.country.languages_spoken, Haitian Creole)[Confidence: 0.7223041653633118] (Nord-Est Department, location.administrative_division.country, Haiti)[Confidence: 0.7090997695922852] ... Question: People from the country that contains Nord-Est Department speak what languages today?</p> <hr/> <p>LLM Response: Based on the given triplets, the country that contains Nord-Est Department is Haiti. The languages spoken in Haiti today are Haitian Creole and French. Therefore, the formatted answers are: ans: Haitian Creole ans: French</p> <hr/> <p>Ground-truth: Haitian Creole French</p>

Figure 4: Correct question-answer example 2.

Question:

What are some of leonardo da vinci most famous inventions?

Retrieved Triples:

Input Triplets:

(Aerial Screw, law.invention.inventor, Leonardo da Vinci)[Confidence: 0.9940758943557739]
(Anemometer, law.invention.inventor, Leonardo da Vinci)[Confidence: 0.9930522441864014]
(Diving suit, law.invention.inventor, Leonardo da Vinci)[Confidence: 0.9929106831550598]
(Triple Barrel Canon, law.invention.inventor, Leonardo da Vinci)[Confidence: 0.9921991229057312]
(33-Barreled Organ, law.invention.inventor, Leonardo da Vinci)[Confidence: 0.9916179776191711]
(Double hull, law.invention.inventor, Leonardo da Vinci)[Confidence: 0.9903085827827454]
(Leonardo da Vinci, law.inventor.inventions, Aerial Screw)[Confidence: 0.9774912595748901]
(Leonardo da Vinci, law.inventor.inventions, Anemometer)[Confidence: 0.9757876396179199]
...

Question: What are some of leonardo da vinci most famous inventions?

LLM Response:

Based on the given triplets, Leonardo da Vinci is credited with inventing several items. Here are some of his most famous inventions listed in the triplets:

ans: Aerial Screw
ans: Anemometer
ans: Diving suit
ans: Triple Barrel Canon
ans: 33-Barreled Organ
ans: Double hull
ans: Armored car
ans: Parachute
ans: Ornithopter
ans: Viola organista

Ground-truth:

Parachute
Diving suit
Aerial Screw
Viola organista
Double hull
33-Barreled Organ
Triple Barrel Canon
Armored car
Ornithopter
Anemometer

Figure 5: Correct question-answer example 3.

<p>WebQSP-WebQTest-801</p> <hr/> <p>Question: What kind of government is egypt?</p> <hr/> <p>Retrieved Triples: Input Triples: (Egypt, location.country.form_of_government, Provisional government)[Confidence: 0.9987523555755615] (Egypt, location.country.form_of_government, Semi-presidential system)[Confidence: 0.9983388185501099] (Provisional government, government.form_of_government.countries, Egypt)[Confidence: 0.9960112571716309] (Semi-presidential system, government.form_of_government.countries, Egypt)[Confidence: 0.9928788542747498] (Egypt, base.aareas.schema.administrative_area.administrative_parent, Earth)[Confidence: 0.00018889721832238138] (Egypt, location.country.currency_used, Egyptian pound)[Confidence: 3.1110161216929555e-05] ...</p> <p>Question: What kind of government is egypt?</p> <hr/> <p>LLM Response: Based on the given triplets, there are two forms of government associated with Egypt: 1. Provisional government 2. Semi-presidential system Therefore, the answers are:</p> <p>ans: Provisional government ans: Semi-presidential system</p> <hr/> <p>Ground-truth: Provisional government Semi-presidential system</p>
--

Figure 6: Correct question-answer example 4.

<p>WebQSP-WebQTest-402</p> <hr/> <p>Question: What does joey jordison play in slipknot?</p> <hr/> <p>Retrieved Triples: Input Triples: (Slipknot, music.musical_group.member, m.010kt51y)[Confidence: 0.6338161826133728] (Slipknot, music.musical_group.member, m.0h2_x53)[Confidence: 0.28905683755874634] (Slipknot, music.musical_group.member, m.01tvxs)[Confidence: 0.12631060183048248] (Slipknot, music.musical_group.member, m.05nn4sx)[Confidence: 0.12411950528621674] (Slipknot, music.musical_group.member, m.0h2_x57)[Confidence: 0.10983577370643616] (m.010kt51y, music.group_membership.group, Slipknot)[Confidence: 0.10371658951044083] (Slipknot, music.musical_group.member, m.05nn4tj)[Confidence: 0.09712643176317215] ...</p> <p>Question: What does joey jordison play in slipknot?</p> <hr/> <p>LLM Response: Joey Jordison is a member of Slipknot and has played various roles in the band. According to the triplets provided, Joey Jordison has played the following roles in Slipknot:Lead guitar,Backing Vocals,Vocals,Bass guitar. Therefore, the answers are:</p> <p>ans: Lead guitar ans: Backing Vocals ans: Vocals ans: Bass guitar</p> <hr/> <p>Ground-truth: Drums</p>
--

Figure 7: Incorrect question-answer example 1.

CWQ-WebQTrn-1259_1997cb4922db71983be26e6a509950f4
<p>Question: Where did the "Country Nation World Tour" concert artist go to college?</p> <hr/> <p>Retrieved Triples: Input Triplets: (Brad Paisley, people.person.place_of_birth, Glen Dale)[Confidence: 0.9620670676231384] (Glen Dale, location.location.people_born_here, Brad Paisley)[Confidence: 0.9496924877166748] (Brad Paisley, music.artist.origin, Glen Dale)[Confidence: 0.6797878742218018] (Country Nation World Tour, music.concert_tour.artist, Brad Paisley)[Confidence: 0.33858561515808105] (Colorado State University, common.topic.notable_types, College/University)[Confidence: 0.3122248351573944] (University of California, Berkeley, common.topic.notable_types, College/University)[Confidence: 0.1940333992242813] ...</p> <p>Question: Where did the "Country Nation World Tour" concert artist go to college?</p> <hr/> <p>LLM Response: Based on the given triplets, we need to find the college Brad Paisley attended. The triplets provide information about Brad Paisley's education, including the institution he attended. From the triplets: - Brad Paisley, people.person.education, m.0h3d7qj - m.0h3d7qj, education.education.degree, Bachelor's degree - m.0h3d7qj, education.education.student, Brad Paisley The triplet indicates that Brad Paisley received a Bachelor's degree, but it does not specify the institution he attended. The identifier 'm.0h3d7qj' does not correspond to a meaningful name of a college or university in the triplets. Therefore, the answer is: ans: not available</p> <hr/> <p>Ground-truth: Belmont University</p>

Figure 8: Incorrect question-answer example 2.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The claims in the abstract and introduction accurately summarize the main contributions, including the introduction of QR-GAT and the overall QSRAG framework, supported by strong empirical results on two KGQA benchmarks.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper discusses limitations such as performance dependence on Top-k selection, potential noise accumulation in large retrieved subgraphs, and sensitivity to confidence thresholding. These are reflected in the further analysis and ablation sections.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include formal theoretical results or proofs; it focuses on model design and empirical evaluation.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper discloses all experimental details, including datasets, model architecture, retriever design, prompt construction, and evaluation metrics. Additional setup such as temperature, Top-k values, and baseline alignment are also documented.

Guidelines:

- The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We plan to release the code and preprocessed datasets upon publication, with instructions for reproducing retrieval and reasoning results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: The paper specifies all key experimental configurations, such as data splits, retriever and LLM types, hyperparameters (e.g., Top-k values), and evaluation setups for each stage.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[No\]](#)

Justification: While the paper provides performance metrics across datasets and variants, statistical significance tests (e.g., confidence intervals, variance/error bars) are not explicitly reported.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: The paper states that experiments were conducted on K100-AI clusters, and specifies that both retrieval and generation used large models such as GLM-4-Air and Llama3.1-8B-Instruct.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research follows the NeurIPS Code of Ethics. It does not involve human subjects or sensitive data, and all used datasets are publicly available.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper briefly discusses that improving structured evidence retrieval helps reduce hallucinations and improve interpretability in LLM reasoning, with implications for safer deployment. Negative societal risks are minimal due to the academic and general-purpose nature of the datasets and models.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The models and datasets used are standard in academic research and do not pose significant misuse risks requiring additional safeguards.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets (WebQSP, CWQ, Freebase) and models (e.g., Llama, GLM-4-Air) are cited properly and used in accordance with their respective licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper proposes a novel retriever architecture (QR-GAT) and QSRAG framework. Documentation and code will be released upon acceptance.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No human participants or crowdsourcing are involved in this work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve any human subjects or user data that would require IRB review.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper clearly states the use of LLMs (e.g., Llama 3.1 8B Instruct, GPT-4o-mini, GLM-4-Air) in the reasoning stage. Their usage is a core part of the proposed methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.